Human control over automation: EU Policy and AI Ethics

Koulu, Riikka

2020-04

Koulu , R 2020 , ' Human control over automation: EU Policy and AI Ethics ' , European Journal of Legal Studies , vol. 12 , no. 1 , pp. 9-46 . https://doi.org/10.2924/EJLS.2019.019

http://hdl.handle.net/10138/318142
https://doi.org/10.2924/EJLS.2019.019

GENERAL ARTICLES

HUMAN CONTROL OVER AUTOMATION: EU POLICY AND AI ETHICS

Riikka Koulu*

In this article I problematize the use of algorithmic decision-making (ADM) applications to automate legal decision-making processes from the perspective of the European Union (EU) policy on trustworthy artificial intelligence (AI). Lately, the use of ADM systems across various fields, ranging from public to private, from criminal justice to credit scoring, has given rise to concerns about the negative consequences that data-driven technologies have in reinforcing and reinterpreting existing societal biases. This development has led to growing demand for ethical AI, often perceived to require human control over automation. By engaging in discussions of human-computer interaction and in post-structural policy analysis, I examine EU policy proposals to address the problematizations of AI through human oversight. I argue that the relevant policy documents do not reflect the results of earlier research which have undeniably demonstrated the shortcomings of human control over automation, which in turn leads to the reproduction of the harmful dichotomy of human versus machine in EU policy. Despite its shortcomings, the emphasis on human oversight reflects broader fears surrounding loss of control, framed as ethical concerns around digital technologies. Critical examination of these fears reveals an inherent connection between human agency and the legitimacy of legal decision-making that socio-legal scholarship needs to address.

Keywords: algorithmic governance, AI ethics, automation, human control, oversight, EU law, legal theory

* Assistant Professor, Director of University of Helsinki Legal Tech Lab, Helsinki.
I would like to thank Jacquelyn Burkell (U Ottawa) for pointing me in the direction of post-structural policy analysis as well as Jörg Pohle (HIIG), Ida Koivisto (Helsinki), Anne Klinefeldter (UNC), and anonymous reviewers for their valuable comments on earlier versions of this article.

TABLE OF CONTENTS

I. INTRODUCTION: HUMAN CONTROL FOR ALGORITHMIC DECISION-MAKING?

Algorithmic decision-making (ADM) systems are used across various fields either to assist and facilitate or to completely automate processes, which previously had mostly been conducted by human decision-makers. Increasing reliance on algorithms, defined as encoded procedures for solving problems by transforming input data into a desired output,[1] is said to contribute to the 'algorithmization' of governance, a distinct form of social ordering that becomes entwined with autonomous algorithm-driven software.[2] Algorithmization has given rise to concerns about the negative

[1] Tarleton Gillespie, 'The Relevance of Algorithms' in Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (MIT Press 2014) 167.

[2] Aneesh Aneesh, 'Global Labor: Algocratic Modes of Organization' (2019) 27(4) Sociological Theory 27(4) 347; Karen Yeung and Martin Lodge, *Algorithmic Regulation* (Oxford University Press 2019).

consequences of data-driven digital technologies, artificial intelligence (AI) and machine learning (ML), terms which are often used interchangeably to refer to the recent phases of the on-going computational turn. In this article, I examine the algorithmization of legal decision-making and the need for AI regulation from a socio-legal perspective.[3] By focusing on how AI use is problematized in the European Union's (EU) emerging AI policy, I explore the problems associated with ADM that law should respond to and the question whether human control over automation is a feasible legislative strategy for addressing these problems. It should be noted that what constitutes a policy problem is not straightforward. Instead, problematizations are created in policy-making.

The emphasis in current algorithm studies has been on algorithmic bias as the most pressing issue related to AI, following the realization that ADM systems reproduce and reinforce existing societal inequalities.[4] In May 2015, an independent news outlet, ProPublica, published an exposé on algorithmic discrimination posed by the presentencing software COMPAS, demonstrating how the system systematically produced higher risk scores for racialized defendants compared to white defendants.[5] Since then,

---

3 Some scholars distinguish between algorithmic and automated decision-making, see e.g. Maja Brkan, 'Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond' (2019) 27(2) International Journal of Law and Information Technology 91, 94. I use these terms interchangeably, as I consider algorithmic decision-making as data-driven automation.

4 Muhammad Ali et al., 'Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes' (2019) arXiv preprint arXiv:1904.02095; Bo Cowgill, 'Bias and Productivity in Humans and Machines' (2019) Upjohn Institute Working Paper 19-309, <https://ssrn.com/abstract=3433737> accessed 27 November 2019; Sara Hajian, Francesco Bonchi, Carlos Castillo, 'Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining' in Balaji Krishnapuram et al (eds), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery 2016) 2125-2126; Sandra G. Mayson, 'Bias in, Bias out' (2019) 128(8) Yale Law Journal 2122; Betsy Anne Williams, Catherine F. Brooks, Yotam Shmargad, 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications' (2018) 8 Journal of Information Policy 78.

5 Jeff Larson, Surya Mattu, Lauren Kircher and Julia Angwin, 'How We Analyzed the COMPAS Recidivism Algorithm' *ProPublica* (23 May 2016)

algorithmic discrimination and other ADM concerns have been widely discussed topics in research as well as in policy-making and the mainstream media.[6] The body of academic literature is rapidly growing and researchers working at the intersections of data science, AI ethics, law and policy studies discuss algorithmic fairness and different means to secure sustainability of ADM systems. The discussions have not emerged out of thin air. For example, computer scientists have long engaged in discussions on what it exactly means for AI systems to be construed as fair.[7]

Against this background, it is somewhat surprising that algorithmization has mostly remained at the margins of socio-legal research.[8] Karen Yeung and

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 15 August 2019. The software's compliance with legal norms have been adjudicated on in Wisconsin Supreme Court's judgment in 2017 in which the court found that the criminal defendant's right to due process was not infringed by the ADM use. See *State vs. Loomis* 881 N.W.2d 749 (2016). See e.g. Liu Han-Wei, Lin Ching-Fu, and Chen Yu-Jie, 'Beyond State v Loomis: Artificial Intelligence, Government Algorithmization and Accountability' (2019) 27(2) International Journal of Law and Information Technology 122. On algorithmic discrimination, e.g. Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, 'Algorithmic Decision Making and the Cost of Fairness' (KDD '17 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2017 797) <https://arxiv.org/abs/1701.08230> accessed 22 August 2019 797–806; Sloane Mona, 'Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice' (Weizenbaum Conference. DEU, 2019).

6 See e.g. Ghaffary Shirin, 'New York City Wants to Make Sure the AI and Algorithms It Uses Aren't Biased. That's Harder Than It Sounds' *Vox* (11 April 2019) <https://www.vox.com/2019/4/11/18300541/new-york-city-algorithms-ai-automated-decision-making-sytems-accountable-predictive-policing> accessed 22 August 2019; Kevin Roose, 'A Machine May Not Take Your Job, but One Could Become Your Boss' *The New York Times* (23 June 2019) <https://www.nytimes.com/2019/06/23/technology/artificial-intelligence-ai-workplace.html> accessed 22 August 2019.

7 See e.g. Ben Hutchinson and Margaret Mitchell, '50 Years of Test (Un)fairness: Lessons for Machine Learning' (Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 2019) <https://arxiv.org/abs/1811.10104> accessed 22 August 2019.

8 ADM in the legal domain is by no means a new phenomenon but instead takes place against the historical backdrop of automation of legal processes through technical

Martin Lodge attribute this underlap of research to doctrinal boundaries that contribute to siloed disciplinary approaches.[9] Some legal scholars have attempted to provide a systematic overview of the ongoing developments. For example, Julie Cohen draws attention to the dynamic reciprocity of technology adoption by noting how law plays a core role in shaping the dynamics of change while being simultaneously restructured in the process.[10] In turn, Mireille Hildebrandt and Katja de Vries emphasize the growing importance of due process and the right to contestation in the face of the computational turn.[11] The socio-legal perspective can be seen as particularly important as it enables us to assess the sufficiency of existing legal and procedural safeguards. The existence of adequate safeguards separates legal decision-making from the many daily decision-making processes now being automated, as the first needs to cater to the overall expectations of coherence, rule of law, and legitimacy of the legal system. Simply put, there is a difference between adequate legal protection when an ADM system is used to curate search engine results compared to automated decisions on refugee status or citizenship, between profiling and decisions with enforceable legal consequences. But in order to assess the existing legal framework of algorithmized governance, we first need to understand what the problems are and what challenges these systems pose. In other words, in the context of

---

systems that has been discussed extensively since the 1950s. Much of the discussion has been framed in terms of AI & Law, although it should be noted that the concept of artificial intelligence (AI) is ambiguous at best. On origins of AI research, see John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955' (2006) 27(4) AI Magazine 12. Also, definitions of AI change over time depending on technological advancements as well as the so-called AI effect, where tasks successfully simulated by machines are no longer deemed AI, see Pamela McCorduck, *Machines who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (A K Peters 2004) 204. For an overview of the development of AI & Law field, see Trevor Brench-Capon, 'A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI & Law', (2012) 20(3) Artificial Intelligence and Law 215.

9 Yeung and Lodge (n 2).

10 Julie Cohen, *Between Truth and Power* (Oxford University Press 2019).

11 Mireille Hildebrandt and Katja de Vries (eds), *Privacy, Due Process and the Computational Turn: The Philosophy of Law meets the Philosophy of Technology* (Routledge 2013).

which concrete concerns are we to evaluate the functioning of law, the effectiveness of existing accountability mechanisms, and the sufficiency of procedural safeguards?

As a response to the public outcry, governments, industry and non-governmental organizations alike are developing ethical frameworks in the hope of enabling fair and trustworthy ADM applications. These AI ethics guidelines provide an opportunity to pose the question above, given that such documents unavoidably need to reflect the perceived problems of AI and to simultaneously construct ethical standards as a solution. In other words, these documents encompass narratives about AI that justify the need for their existence. Sometimes framed as 'ethics-washing', the instruments have been criticized for their non-binding nature, the lack of a clear scope of application, and limited interpretative advice of fairness for programmers and administrators of justice alike, all of which contributes to their limited ability to regulate the development and use of AI systems.[12] In terms of legal sources, these instruments can be described as soft law[13] that lack formal validity but influence how policy issues are perceived. Not all soft law instruments are alike, however; instruments created by powerful supranational institutions such as the EU also rely on the authority of the institutions and not simply on the strength of their arguments. In this sense, soft law may also foster the creation of hard law by providing early conceptualizations of relevant AI policy issues that allegedly need to be addressed. The AI ethics guidelines usually advocate human oversight as a meaningful protection against the negative consequences of technology use.

---

[12] See e.g. Thilo Hagendorff, 'The Ethics of AI Ethics: An Evaluation of Guidelines' (2019) <https://arxiv.org/abs/1903.03425> accessed 22 August 2019; Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) Big Data & Society 1; Daniel Greene, Anna Lauren Hoffmann, and Luke Stark, 'Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning' (Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019) DOI: 10.24251/HICSS.2019.258 accessed 22 August 2019.

[13] See e.g. Alan Boyle, 'Some reflections on the Relationship of Treaties and Soft law' (1999) 48(4) International and Comparative Law Quarterly 901, 901-913. According to Boyle, soft law is defined by its non-binding nature, focus on general principles instead of rules, and lack of direct enforceability.

Does this mean that hard law regulation should also aim to include human control as a legal protection?

This article does not repeat the critique against AI ethics in policy-making, although the established shortcomings do form its starting point. I discuss one solution proposed in the EU's policy-making, namely human control, referred to as Human-in-the-loop (HITL), human oversight or intervention, human-on-the-loop (HOTL), or human-in-command (HIC).[14] There is also a terminological connection between human control and the so-called human-centric approach, which also poses a similar linguistic focus on human agency.

Answering the question on legislative strategy for ADM requires us to assess the feasibility of human control from a socio-legal perspective, particularly as the EU is now developing the structures and processes to govern ADM systems, which are then established as legal rights, obligations, and safeguards. Political choices on regulatory objectives are translated into legal concepts and thus operationalized within the legal system. Once employed, these objectives and regulatory choices can become embedded within the legal structures and cannot be fundamentally contested. Human oversight may become a central procedural mechanism for automated decisions, but once it defines procedural rights and obligations it is more difficult to present a fundamental critique of its feasibility. That is the reason why it is important to ask now whether human control can fulfil its promise, requiring us to consider the problems of AI that call for human control. A regulatory strategy built on false beliefs about the strengths of human control may fail to provide adequate legal protection for those subjected to automated legal decision-making.

---

[14] European Commission, Independent High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> accessed 22 August 2019 (Hereinafter Guidelines), 16; see also Communication COM(2019) 168 final from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions on Building Trust in Human-Centric Artificial Intelligence [2019] <https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF> accessed 22 August 2019, 4.

At first glance, human control seems like a plausible solution, as ultimately it aligns with law's anthropocentricity, reflected in the fact that law recognizes only human actors as objects of regulation, not machines. At times, law goes to great lengths to uphold at least the fiction of this anthropocentricity, for example by granting legal personhood to corporations and non-human organizations. Hence, it is perhaps not surprising that the importance of keeping humans in control of automation is widely agreed upon in legal scholarship.[15] The reasons given may be instrumental, such as of the need to allocate responsibility to human actors due to legal liability regimes.[16] However, there seems to be a more fundamental argument that considers the human element as being intrinsically indispensable, although this is not elaborated on in great length. Instead, human agency, participation and control are portrayed as uncontestable necessities that are ultimately connected with democratic legitimacy. For example, John Danaher contends that '[l]egitimate decision-making procedures must allow for human participation in and comprehension of those decision-making procedures' and that, because reliance on ADM limits active human participation, the systems impose a fundamental threat to legitimacy that he considers difficult to accommodate or resist.[17] In her work on the intersections of law, technology and philosophy, Mireille Hildebrandt addresses similar issues of justification and discusses the need for protection of 'what is uncountable, incalculable or incomputable about individual persons', which comes under threat in the context of automated decision-making, where contestation by those subjected to automation plays a vital role.[18] In contrast, others focus on

---

15 See e.g. Woodrow Hartzog, 'On Questioning Automation' (2017) 48 Cumberland Law Review 1; Michael Schmitt and Jeffrey Thurnher, 'Out of the loop: autonomous weapon systems and the law of armed conflict' (2012) 4 Harvard National Security Journal 231; Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 Washington Law Review 1.

16 Madeleine Elish, 'Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction' (2019) 5 Engaging Science, Technology, and Society 40, 41.

17 John Danaher, 'The Threat of Algocracy: Reality, Resistance and Accommodation' (2016) 29(3) Philosophy & Technology 245, 254.

18 Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20(1) Theoretical Inquiries in Law 83, 83-121.

the fabricated and performative nature of human intervention. For example, Sheila Jasanoff draws attention to the 'human pretensions of control over technological systems', which demands for critical re-examination.[19]

This article is built on Jasanoff's call for critical examination of the feasibility of human control over automation. I argue that the focus on human control in policy decisions over automation is insufficient and misguided. I build this argument in two steps. In section II, I discuss the origins and limitations of human control over automation, considering research conducted on human-computer interaction (HCI). The HCI literature provides insight into the potential and shortcomings of human control and hence explains the situations and conditions in which human control may be worthwhile. This HCI perspective is often missing in both policy-making as well as in socio-legal scholarship. In section III, I engage in critical analysis of the EU's three policy documents on AI in order to identify the situations in which the documents advocate for human control as a meaningful precaution. While the explicitly expressed problems do form a starting point for this analysis, they do not provide an exhaustive overview. Instead, policy documents come embedded with implicit assumptions about the problems they aim to target and these problematizations are not necessarily the same as those explicated. The acknowledgment of how AI is problematized both explicitly and implicitly is necessary to assess the feasibility of human control for legal protection. By engaging in post-structural policy analysis, described in further detail in section II.2, I aim to reveal the implicit assumptions behind the chosen approach to human control. By contrasting the explicit and implicit problematizations, we can provide a more nuanced understanding of the perceived problems with AI that human control is thought to address. Finally, in section IV, I return to the discussion of human control as a regulatory strategy and the role of human agency in the legitimacy of decision-making. By incorporating perspectives from legal theory and European law, HCI and policy analysis, I hope to combine theoretical assessment of the feasibility of human control with a close reading of policy documents and interconnect these with *de lege ferenda* discussions surrounding AI. This approach contributes to a more comprehensive socio-

[19] Sheila Jasanoff, 'Technologies of Humility: Citizen Participation in Governing Science' (2003) 41(3) Minerva 223.

legal overview of the object of AI regulation and the complex interconnections between law, technology, and policy.

II. Origins and Limitations of Human Control

1. Human-Machine Interaction Research

In this section, I discuss some of the early research in human-computer interaction in order to demonstrate the context in which the early formulations of human control over automation emerged. Such genealogical analysis is necessary in order to understand what it means to establish human control as the core means of organizing the division of labor, as well as legal liability, between the human decision-maker and the ADM system. Human control over automation comes with conceptual baggage related to its early context and the level of technological development at the time that we need to understand in order to critically examine its feasibility as a potential regulatory strategy.[20] The early work on human control was built on perception that humans and machines have differing capabilities, requiring a separation between the tasks entrusted to humans and machines respectively. This human/machine dichotomy has since been challenged by research focused on collaboration rather than division, but it still provides an important framing. In policy documents, the historical context of human control is typically not elaborated, meaning that the underlying assumptions remain outside the scope of policy debate.

Originally framed in terms of human-in-the-loop rather than human control, early iterations are often traced back to post-war work on aviation security.[21] Some of the early iterations of human-in-the-loop were developed in human factors research in relation to aviation security in the US, with the objective of reducing human error and enhancing safety through a focus on the interaction between humans and computers. The early work on human factors was interested in function allocation, i.e. which tasks should be

[20] On conceptual baggage of key concepts see, e.g. Jan Ifversen, 'About Key Concepts and How to Study Them' (2011) 6(1) Contributions to the History of Concepts 65, 73.

[21] See e.g. Elish (n 16) 40–60.

automated by computers and which ones should remain within human control.

A starting point for this line of inquiry can be traced back to 1951, which saw the publication of the so-called Fitts list, which was meant to provide background information for policy makers. The list was drafted by Paul Fitts, a former US Air Force Lieutenant Colonel and psychologist at the University of Ohio, who went on to develop a mathematical model to predict human motion called Fitts's law. The so-called HABA-MABA model ('humans are better at, machines are better at') included 11 statements to describe tasks humans are better at accomplishing and which are more easily performed by machines. According to Fitts, humans surpass machines in cognitively challenging tasks such as perception, judgment, improvisation, and long-term memory, whereas machines are better than humans in tasks that require speed, power, computation, replication, simultaneous operations, and short-term memory.[22] Fitts list has remained a seminal work of function allocation research and, as will be examined in further detail in section III, the foundational assumptions have later been adopted and expanded in broader discussions on the necessity of human control over technology, most recently in AI ethics discussions.[23]

While the HABA-MABA model now seems somewhat outdated, in its time it provided an adequate description of which tasks could be automated. The model reflected the contemporary state-of-the-art of technological development. In addition to technological progress, early iterations of human control also reflected political and ideological choices of the time, as

---

[22] Paul M. Fitts (ed), *Human Engineering for an Effective Air-Navigation and Traffic-Control Aystem* (National Research Council, Division of Anthropology and Psychology, Committee on Aviation Psychology 1951).

[23] See e.g. Joost de Winter and Dodou Dimitra, 'Why the Fitts List has Persisted Throughout the History of Function Allocation' (2014) 16(1) Cognition, Technology & Work 104. On criticism, see Meg Leta Jones, 'The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles' (2015) 18 Vanderbilt Journal of Entertainment & Technology Law 77, 106: 'The Fitts List has been heavily criticized as an intrinsically flawed descriptive list, little more than a useful starting point, insufficient, outdated, static, and incapable of acknowledging the organizational context and complementary nature of humans and machines'.

the model was adapted to the highly politicized topic of space travel in the Apollo program in 1960-1972. In space aviation, the involvement of a human operator was also considered necessary for automated operations with the concession that inclusion could take place remotely.[24] Due to the geopolitical dimension, organization of human control became a question of ideological choice. David Mindell describes how preference given to human control reintroduced the perceived political differences between the American and Soviet approaches.[25] Interestingly, the HABA-MABA model still persists as a key conceptualization of human-machine interaction and, as such, is often referred to in legal discussions on automation, albeit often critically.[26] In the 1980s, the human/machine dichotomy was increasingly superseded by the notion of 'human-centered design',[27] although it is unlikely that the latter concept actually signified separation from the earlier doctrine, despite the terminological shift. Simply put, if everyone still refers to the HABA-MABA model, even critically, the model still persists as the locus of discussions on the central framing of human-computer interaction, and consequently continues to frame considerations concerning potential solutions.

Function allocation research and later work on teleoperations, human-machine interaction, and cognitive engineering have demonstrated some of

---

[24] 'One of the pre-requisites for taking the man out of the systems operation must be the capability to describe very carefully, and in some detail, the characteristics of the operation before it starts. Of course, in some instances the man can be included by leaving him on the ground and providing him with necessary intelligence'. See Richard Horner, 'Banquet Address before the first Annual Awards Banquet of the Society of Experimental Test Pilots' (1957) 2(1) SETP Quarterly Review 1, 7, as referenced in David Mindell, *Digital Apollo: Human and Machine in Space Flight* (MIT Press 2008) 19.

[25] 'Keeping the astronauts "in the loop," overtly and visibly in command with their hands on a stick, was no simple matter of machismo and professional dignity (though it was that too). It was a well-articulated technical philosophy. It was also necessary to achieve the political goals of the space program and show that the classical American hero—skilled, courageous, self-reliant—had a role to play in a world increasingly dominated by impersonal technological systems (especially in contrast to the supposedly over-automated Soviet enemy)'. See Mindell (n 24) 5.

[26] See Jones (n 23) 130. In fact, citations on the Fitts list have steadily increased during the last decades. See De Winter and Dodou (n 23) 2.

[27] Jones (n 23) 112.

the inherent shortcomings of human control over automation. For various reasons, from boredom at routine monitoring to automation bias and alert fatigue, humans generally perform badly as supervisors of automated technical systems.[28] These 'ironies of automation' were discussed in 1983 by Lisanne Bainbridge, who explained how automation design 'still leaves the operator to do the tasks which the designer cannot think how to automate', despite the intention to replace human control. These tasks usually include monitoring and take-over functions which humans have been shown to perform badly.[29] Bainbridge argues that 'by taking the easy part of his task, automation can make the difficult parts of the human operator's task more difficult'.[30] Similarly, the notion that accidents related to technical systems follow from human error was contested by sociologist John Perrow, according to whom systemic or 'normal accidents' follow from combined effects of tightly coupled complex systems that have high risk potential.[31] Thus, accidents are unavoidable in the sense that they cannot be prevented by simple design choices. In light of technological development and the introduction of the relatively autonomous ADM systems currently in use, the recent HCI research discussed above seems to provide a better account of the limitations of human control than the human/machine dichotomy. Based on these insights, the scope for human control over automation seems relatively narrow in practice.

---

28 'There is much evidence that people are not good monitors of automation' for various reasons, including boredom that ensues from monotonous tasks, see Thomas B. Sheridan, Skaar S. B., Ruoff C. F., 'Human Enhancement and Limitation in Teleoperation' (1994) 161 Progress in Astronautics and Aeronautics 43, 54; Elish (n 16) 50 'skills atrophy when automation takes over'; on alert fatigue in the medical field, see Rush Jess et al., 'Improving Patient Safety by Combating Alert Fatigue' (2016) 8(4) Journal Graduate Medical Education 620, 620–621.

29 Lisa Bainbridge, 'Ironies of Automation' (1983) 19(6) Automatica 775, 775-779.

30 Interestingly, she considers human oversight as a necessity for complex automation: 'There will always be a substantial human involvement with automated systems, because criteria other than efficiency are involved, e.g. when the cost of automating some modes of operation is not justified by the value of the product, or because the public will not accept high-risk systems with no human component'. Ibid 777.

31 John Perrow, *Normal Accidents: Living with High-Risk Technologies* (Basic Books 1984).

One might expect that AI policy would be informed by these observations. Instead, it seems that the early human/machine dichotomy is still reproduced in policy-making without including later critical appraisals. Hence, policy documents portray human control in opposition to unstoppable technological change, rather than as hybridization of complex socio-technical systems, i.e. seamless collaboration between humans and artificial systems. In their critical analysis of AI ethics documents, Greene et al. point out that

> the precise reasons why AI/ML are matters of ethical concern differ from organisation to organisation. Some lean on the language of distributive justice, arguing AI/ML's benefits and penalties will be unevenly distributed.[32]

Greene et al. argue that AI ethics guidelines reflect ethical universalism and determinism, which means that ethical concerns are seen as a universal, cross-species force of nature to which humans can only react. Simultaneously, human agency is advocated as a plausible solution, although in the form of expert oversight instead of public mass movement. Jones draws attention to the arbitrariness of policy-making that operates on the logic of human oversight: 'when presented with an automation-related problem, law and policy responses have been to preserve or protect an explicit value by simply inserting or removing a human from the loop, which actually ends up backfiring'.[33]

Furthermore, Madeleine Clare Elish suggests, human oversight may be used detrimentally to assign guilt and responsibility to humans. Elish introduces the concept of 'a moral crumple zone to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system'.[34] Drawing on investigations of the Three Mile Island nuclear accident in 1979 and the Air France Flight 447 crash in 2009, Elish demonstrates how in these two cases, the accidents were attributed to human error despite the fact that both resulted from a complex set of factors related to human-machine interaction, as well as to system design. According to Elish, law and policy play a role in the creation of moral crumple zones, as attribution of liability in aviation

---

32 Greene et al. (n 12) 2127.

33 Jones (n 23) 81.

34 Elish (n 16) 40.

demonstrates: certification standards recognize only mechanical failure to give rise to accountability and hence only a human pilot can be the source of malfunction in situations of shared control.[35]

Limitations and problems of human oversight are widely acknowledged in research, leading to efforts to improve the inherently flawed human-facing control of automation. For example, Brkan discusses the minimum acceptable level for meaningful human oversight in light of EU legislation, thus addressing the issue of 'rubber stamping', when human control becomes mostly performative.[36] Drawing from the research field of AI & Law and 'by design' approaches, Almada proposes reinterpretation of human intervention in a manner that would complement post hoc oversight with an ex ante approach he calls 'contestability by design', through which the safeguards and data of the subject's rights stipulated in article 22 of the General Data Protection Regulation (GDPR)[37] would be embedded in the technical design of the ADM system.[38] In turn, from the computer science perspective, Sirajum et al. argue that HITL should be a central system design principle, requiring solutions to certain challenges, most important of which is to determine 'how to incorporate human behavior models into the formal methodology of feedback control'.[39]

---

35 Ibid 50.

36 See e.g. Brkan (n 3), where she contends that rubber stamping is not enough for meaningful intervention necessitated by GDPR article 22 but instead the overseer needs to possess authority and capability to change the decision.

37 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

38 Marco Almada, 'Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems' (International Conference on Artificial Intelligence and Law ICAIL'19, 2019) <https://www.researchgate.net/profile/Marco_Almada/publication/327602212_Human_intervention_in_automated_decision-making_Toward_the_construction_of_contestable_systems/links/5cc64eb8a6fdcc1d49b76103/Human-intervention-in-automated-decision-making-Toward-the-construction-of-contestable-systems.pdf> accessed 23 August 2019.

39 Sirajum Munir et al., 'Cyber Physical System Challenges for Human-in-the-Loop Control' (8th International Workshop on Feedback Computing, 2013)

Others have provided alternative problematizations of, while still advocating some form of human control as a potential solution. For example, Liu et al. attribute the problem partly to the current focus on the technical perspective of AI development that disguises the embedded heterogenous power relations.[40] Rahwan assigns the problem to a lack of societal commitment, which could be resolved by 'looping in' society.[41] Within the Human-AI Interaction field, Amershi et al. identify the core problem as being the unpredictability of AI-infused systems, which results from uncertainty and leads to false positives and false negatives; the remedy lies, they suggest, in improving user interface design following generally accepted design guidelines.[42]

In summary, decades of research on human-machine interaction have developed nuanced approaches to human control and simultaneously demonstrated its practical limitations over automated systems. But does policy-making reflect these insights? And if not, do AI ethics guidelines end up reproducing these 'human pretensions of control over technological systems'?[43] Do the AI policy documents take it for granted that human control ensures adequate ethical and legal safeguards?

2. Engaging in Post-Structural Policy Analysis

The fact that human control is advocated in AI policy, despite the limitations established by HCI, suggests that either the policy-making is built on false assumptions about the potential of such control or, alternatively, that the emphasis on human control serves a purpose other than *de facto* oversight. As discussed in section I, this purpose might involve the justification and overall legitimacy of decision-making, as some legal scholars suggest. But in order to

<https://www.usenix.org/system/files/conference/feedbackcomputing13/feedback13-munir.pdf> accessed 23 August 2019.

40 Liu et al. (n 5).

41 Iyad Rahwan, 'Society-in-the-Loop: Programming the Algorithmic Social Contract' (2018) 20(1) Ethics and Information Technology 5, 7.

42 Saleema Amershi et al, 'Guidelines for Human-AI Interaction' (Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, 2019) <https://dl.acm.org/citation.cfm?id=3300233> accessed 23 August 2019.

43 Jasanoff (n 19).

substantiate this claim, we need to look closer at the policy documents to identify the problems which human control is considered to address.

What does a problem description in a policy document entail? I proceed from the observation that problematizations are fabricated in the course of policy-making. In this sense, problem representations in policy-making are not neutral. Instead, linguistic choices reflect the power to decide which issues are worthy of policy action and which issues are not. This perspective aligns with the argument presented by Liu et al., namely that the narrow focus on technology diverts attention from the heterogenous power relations of AI development.[44] In a similar vein, I argue that AI policy documents include, in addition to the problems they explicitly point to, implicit assumptions about the problems underlying the proposed solution of human control. To understand the intricacies of problem presentations better, I have employed a Foucault-influenced post-structural policy analysis called the 'What's the Problem Represented to Be?' or WPR approach in order to reveal how human control reinforces the old distinction between human and machine and attributes legitimacy creation only to human agency.

What is the added value of this focus on problematizations for AI policy analysis? Foucauldian sociology has been particularly interested in the intricate ways in which power works through language, which often remains beyond the scope of more socio-legal approaches to policy analysis. Although such analysis might seem merely descriptive from the legal viewpoint, the objective of Foucauldian policy analysis is in fact diagnostic. As Nikolas Rose puts it, analytics of governmentality 'seek an open and critical relation to strategies for governing, attentive to their presuppositions, their assumptions, their exclusions, their naivities and their knaveries, their regimes of vision and their spots of blindness'.[45] Consequently, the focus on problematizations aims to broaden the space of possible policy solutions. Hence, this diagnostic examination serves the needs of the socio-legal perspective as it provides a deeper understanding to support informed policy decisions on *de lege ferenda*. Other socio-legal scholars have conducted similar analyses in other fields of law. For example, Dent applies Foucauldian analysis

---

44 Liu et al (n 5).

45 Nikolas Rose, *Powers of Freedom: Reframing Political Thought* (Cambridge University Press 1999) 9.

to examine the copyright regime as a governmentalist practice diffused throughout society.[46] At the core of governing lies the process of subjectification, how subjecthood is not naturally given but created through governing practices that contribute to the process in which human persons turn themselves into subjects of governing. He argues that the field cannot be characterized by any single problematization but rather is filled with different government rationalities that can be made visible through a complete genealogical examination of copyright practices. In his analysis, 'problematization is both a process of governance and a technique for investigating the acts of governing'.[47] To this end, he argues, the advantage of problematizations is that it enables us to perceive multiple rationalities and purposes instead of a static 'monolithic, ahistorical problematization of (self-) expression'.[48]

Similarly based on the Foucauldian sociology of problematizations, Carol Bacchi has explored how problems are constituted in policy documents and how governance is organized through these problematizations, with the objective of exposing how the political agenda behind 'chosen' problems insidiously defines what is possible or impossible to ask, which outcomes are desired or undesired, which perspectives are included and which excluded – in short, what the policy debate is about.[49] The WPR approach contests 'the common view that the role of governments is to solve problems that sit outside them, waiting to be "addressed"' and provides step-by-step

---

46 Chris Dent, 'Copyright, Governmentality and Problematisation: An Exploration' (2009) 18(1) Griffith Law Review 129, 131.

47 Ibid 133.

48 Ibid 141.

49 'To say that policies *create* "problems" as particular sorts of problems, does not mean to suggest that governments set out to *produce* homelessness or poverty, or even to deliberately represent homelessness or poverty in particular ways. Rather, the proposition is that the specific policy or policy proposal contains *within it* an implicit representation of the 'problem', referred to as a problem representation. This proposition relies upon a simple idea: That what we propose to do about something indicates what we think needs to change and hence what we think is problematic – that is, what the 'problem' is represented or constituted to be'. See, Carol Bacchi, 'Problematizations in Health Policy: Questioning How 'Problems' Are Constituted in Policies' (2016) SAGE open <https://doi.org/10.1177/2158244016653986> accessed 23 August 2019.

instructions for elaborating how problems are made within policy-making practices.[50] The analysis involves 'working backwards' from proposed solutions to problem representations and, following a set of questions, drawing attention to the underlying presuppositions and assumptions as well as the emergence and effects of the said problem representation. In the context of the present paper, this means working backwards from the proposed solution of human control over automation to question what are construed as the 'problems' of ADM systems and AI in general. For my analysis, this means looking at how human control is formulated in the policy documents in order to reveal the embedded assumptions the solution presupposes. What does human control tell us about the nature of AI problems in the EU's policy? What history, context, and narrative are generated in these policy documents? Do the policy documents reflect a reasonable understanding of the possibilities and limitations of human control as they are discussed in HCI research?

The WPR approach lists potential questions that guide the critical analysis, starting by identifying problem representations in search of 'a way to open up for questioning something that appears natural and obvious'.[51] I focus in particular on questions that aim to reveal the hidden ontological assumptions behind policy formulations and what is left unsaid (and thus excluded from discussion) by these formulations: what deep-seated presuppositions or assumptions underlie this representation of the 'problem'? What is left unproblematic in this problem representations? Where are the silences? Can the 'problem' be conceptualized differently?[52] The last step in Bacchi's approach is self-problematization, the reflexive application of the critical approach to the analyzer's own argumentation to reveal the selective choices that motivate it. In line with this approach, I argue that human control as a

---

50 Carol Bacchi and Susan Goodwin, *Poststructural Policy Analysis: A Guide to Practice* (Palgrave 2016) 14. WPR approach is not interested in 'how different *people* might problematize the issue but how the *policy itself* problematizes it' (p. 17). Hence, the focus is on how problematizations are created by policy *itself*, not how individuals and organizations involved in policy-making processes perceive them. Complex policy documents often contain more than one problem presentation (p. 20), as is also the case with the EU's AI ethics documents.

51 Ibid 20.

52 Ibid 20-21.

solution to AI problems is built on a premise not unlike that of human/machine dichotomy of HABA-MABA model, namely that the actions of humans and technological systems can be clearly distinguished from one another and the former put in charge of the latter. The policy documents ultimately build a hopeful narrative: AI risks are construed as potentially harmful for human autonomy, but with human control these harms can effectively be prevented. Although aspirational, the narrative does not necessarily hold true in light of HCI research.

III. Making the Implicit Explicit: AI Problematizations in EU Policy

1. The Explicit Objectives of EU Policy: Putting People at the Center of AI Development

In this section, I analyze three documents that reflect the EU's current policy on AI ethics. The first of these documents is the Commission's communication on Artificial Intelligence for Europe from spring 2018 ('the Strategy'), mandated by the Council, which establishes the need for a European approach in order to reap the advantages of AI.[53] The second document is the Ethics Guidelines for Trustworthy AI ('AI HLEG') drafted by the Independent High-Level Expert Group set up by the Commission.[54] The expert group delivered its first draft in December 2018 and, after stakeholder consultation, a revised version in April 2019.[55] The third document is the Commission's communication in April 2019 on Building Trust in Human-Centric Artificial Intelligence ('Communication') that incorporates the key points of the AI HLEG guidelines.[56] I first discuss how AI policy issues are framed and positioned in these documents, considering

---

53 Communication COM(2018) 237 final from the Commission to the European Parliament, the European Council, The Council and the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe [2018] <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF> accessed 22 August 2019, 2.

54 Ibid section 3.3.

55 See COM(2019) 168 final (n 14). As soft law, the Guidelines are meant to be adopted by stakeholders on a voluntarily basis.

56 Ibid.

the terminological ambiguity of AI and what the perceived relationship between law and ethics is. In addition, I examine the intended usage and form as well as explicit expectations linked with human control. In section III.2, I then locate what has been left unsaid in the hope of finding out what remains beyond the scope of these policy initiatives.

To understand better the explicit problems of AI these documents aim to address, we first need to look into what is meant by AI, i.e. from which qualities do perceived problems emerge. Interestingly, AI is not defined in technological terms in any of the documents but instead by reference to digital transformation and the increasing autonomy of AI systems. According to the Strategy, AI is one of the most strategic technologies of the 21st century and is transforming the world, society, and industry like the steam engine and electricity in the past. AI is defined as 'systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals'.[57] The systems are both software-based and embedded in hardware and often require data to improve their performance. Hence, AI is seen to refer to relatively autonomous algorithmic models that infer outputs from input data. In short, AI is perceived as a combination of relatively autonomous data-driven technologies. This conception of AI is unsurprising given that data governance is at the core of EU technology policy. Furthermore, the GDPR creates a normative basis for automated decision-making that connects data subject's legal protection with human intervention. Article 22(1) of the GDPR provides for the right for a data subject not to be subjected to a decision based solely on automated data processing. Although exceptions to the main rule are stated in article 22(2), these may only be applied with suitable measures for the data subject's legal protection, the minimum standard stated in 22(3) being the data subject's 'right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision'.[58] The existing regulation also forms the basis for the development of an AI-specific framework around human intervention.

What then are the stated objectives of the policy documents? The Communication states that the aim of the emerging AI ethics regime is to

---

57 COM(2018) 237 final (n 53) 2.

58 On GDPR article 22, see e.g. Brkan (n 3).

place people at the center of the development of AI – hence the formulation, 'human-centric AI'.[59] In the Strategy, the goals are described somewhat differently, in terms of boosting the EU's technological and industrial capacity, preparing for socio-economic changes brought by AI, and ensuring an appropriate ethical and legal framework.[60] Understandably, the focus of the policy actions is on economic measures, given the EU's legislative mandate. The Strategy identifies the lack of trust and accountability as key AI-related problems. The new opportunities generated by AI are contrasted with the possible uses of AI for 'malicious ends', whereas the challenges and risks are located in the 'areas of safety and liability, security (criminal use or attacks), bias and discrimination'.[61] The proposed solution is to develop AI ethics guidelines in collaboration with all stakeholders following the European Council's original mandate.

The Strategy connects the ethical standards with the EU Charter of Fundamental Rights and the values listed in article 2 of the Treaty on European Union: respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. The Guidelines reflect the problems identified in the Commission's Strategy, i.e. the lack of trust and accountability, suggesting people's mistrust would prevent the adoption of AI: 'without AI systems – and the human beings behind them – being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered, preventing the realization of the potentially vast social and economic benefits that they can bring'.[62] It is thus claimed that the crucial problem related to AI is that applications would not be used, a problem that can be solved by increasing trustworthiness. Trustworthiness, in turn, is seen to be achieved by a combination of legal compliance, ethical principles, and technical and social robustness.[63] The Guidelines do perceive that AI

59 COM(2019) 168 final (n 14) 1.

60 COM(2018) 237 final (n 53) 4.

61 Ibid.

62 Guidelines (n 14) 4-5.

63 Ibid 5.

applications also present other risks, although these are not further elaborated on.[64]

Ultimately, it is human autonomy that is seen to be threatened. Human autonomy is constituted as the protected good and the core ethical principle that necessitate human oversight as a safeguard:

> The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings. Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems. AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work.[65]

This means that the perceived threat of AI is the loss of human autonomy, particularly of those humans who find themselves interacting with AI systems. The proposed solutions are human-centric design and opportunities for human choice, which can be realized through human oversight of AI systems. This proposal implies that human control is all that is needed to ensure ethical AI systems. However, this is contestable given the insights from HCI research, which demonstrated the limited capabilities of humans as overseers of automated systems. Simply put, human control over automation often fails in practice. Furthermore, there is a more fundamental challenge to this approach: imposing responsibility for ethical AI on the human controller might be unreasonable from the perspective of the controller's legal protection.

Human control as the solution to AI problems becomes visible particularly in the HLEG Guidelines, which present oversight as being necessary to

---

64 'While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately', see ibid 4.

65 Ibid 12.

ensure that 'an AI system does not undermine human autonomy or cause other adverse effects'.[66] This formulation explicates the problem as follows: without such oversight, the machines may be detrimental to human self-determination. Human control seemingly does not need to be comprehensive in order to be considered effective protection. The Guidelines explain different degrees of human oversight from step-by-step monitoring in the form of human-in-the-loop to overall monitoring of human-in-command, noting that lower levels of human oversight should be accompanied by other safeguards:

> Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.[67]

Finally, the Guidelines bring forward a concrete assessment list targeted towards AI practitioners, with questions meant to ensure adequate human oversight. The list includes a set of questions on the level of human involvement, identification of the human overseer and moments and tools for intervention, existence of detection and response mechanisms for

---

66 Ibid 16.

67 Ibid.

autonomous AI systems, and the inclusion of a stop button or procedure for safely aborting an operation.[68]

The expert group's recommendations on human oversight were included almost word for word in the Commission's 2019 communication, 'Building Trust in Human-Centric Artificial Intelligence'. In addition to formatting and slight changes of wording, the Communication connects human control measures with the adaptability, accuracy and explainability of AI-based systems, which in the Guidelines were discussed in terms of technical and social robustness and data governance.[69] The Communication also contains a preamble to the description of human oversight as a key requirement for trustworthy AI. Finally, the user's overall wellbeing is highlighted as being central to the system's functionality.[70] The Communication also sets out the next steps in establishing an AI framework, which include, inter alia, stakeholder feedback on the feasibility of the assessment list and potential revisions, as well as building the EU's leadership role in international policy settings with the objective of creating a related assessment mechanism. In addition, more funding will be targeted to research on explainability and advanced human-machine interaction.

In summary, the EU's policy documents on AI ethics create high expectations that human oversight will safeguard human autonomy in the development and use of AI applications. The Commission's Strategy considers the lack of trust and accountability as the main concerns associated with AI, whereas the expert group's Guidelines present the undermining of human autonomy as one of the main problems that can be remedied by human oversight. The Communication of 2019 repeats these concerns and, unlike the earlier documents, connects required human control with accuracy and explainability. Interestingly, the question of interaction between humans and machines is only mentioned in the conclusions of this most recent policy document. Based on these observations, the policy documents reflect the assumption that human control constitutes an

---

68 Ibid 27.

69 Ibid 17-18.

70 COM(2019) 168 final (n 14) 4.

effective accountability mechanism for the protection of human autonomy in the face of increasing AI use.

2. Discovering the Implicit: Human Autonomy as the Last Stand against the AI Tidal Wave

As discussed, post-structural policy analysis is particularly interested in locating the silences created in policy-making, as these frame the scope of what is construed as possible policy action. Policy documents typically aim to justify legislative action by imposing them as the right solutions to identified problematizations, but these stances are ultimately opinions about what needs to be fixed.[71] How does this perspective play out with the EU's AI policy? What is left unsaid? I address these questions by working backwards from the proposed solution of human oversight, with the specific objective to follow to which actions and by whom the human control is extended. Who is the object of human oversight?

The Commission's Strategy on Artificial Intelligence for Europe recognized the need to 'ensure an appropriate ethical and legal framework', suggesting that the current framework is not sufficient to ensure trust and accountability. In other words, the legal and ethical framework needs fixing. What, then, are the proposed solutions? The Commission's proposal is to carry out as soon as possible the Commission's agenda as defined in an earlier policy document, the Digital Single Market Strategy, including measures such as enabling free flow of non-personal data 'that will be a key enabler for the development of AI'.[72] At the same time, the quick adoption of these measures is 'essential as citizens and businesses alike need to be able to trust the technology that they interact with, have a predictable legal environment and rely on effective safeguards protecting fundamental rights and freedoms'.[73]

These statements reveal a two-sided and inherently conflicting perception of AI: on the one hand, AI is a good thing and needs to be actively enabled by regulatory measures; on the other, AI threatens fundamental rights and

---

71 See n 49.

72 COM(2018) 237 final (n 53) section 3.3.

73 Ibid.

therefore needs to be reined in with legal and ethical guidance. The general public's trust, in turn, is linked with its understanding about the technical underpinnings of AI systems and thus explainability of the system is a key measure for solving AI-related problems:

> To further strengthen trust, people also need to understand how the technology works, hence the importance of research into the *explainability of AI systems*. Indeed, in order to increase transparency and minimize the risk of bias or error, AI systems should be developed in a manner which allows humans to understand (the basis) of their actions.[74]

This statement is built on the assumptions that, firstly, humans are indeed capable of understanding complex technical systems and, secondly, that the human activities of seeing and understanding are enough to protect those values that are threatened by AI. Read this way, the problem with AI is also about people's lack of understanding that can be solved by human oversight which addresses this understanding.

These risks are portrayed as unavoidable characteristics of the current technologies and thus constructed as 'normal' AI-related problems. By the use of passive language, these problems are attributed to the technology itself, not to the software developers and system architects, to institutional practices or organizational and market structures. In the Strategy, legal and ethical concerns are addressed by product liability, data protection, cybersecurity and intellectual property rights. At the same time, other legal mechanisms like competition law, administrative and procedural law, as well as tax law are excluded from examination, although these fields do provide effective mechanisms to ensure legal protection. Competition law in particular could be considered, because regulation of markets is a powerful tool that can also be used to guide commercial AI development.[75] In light of these three AI policy documents, AI problems, it seems, are problems created by the technology, not by humans, and these problems should primarily be addressed by product liability and data protection regimes, i.e. only certain areas of law. This focus on certain legal fields frames socio-legal

---

74 Ibid.

75 Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2016).

problem solving, pushing us to address AI issues primarily within these legal regimes, which then diverts attention from alternative legal remedies.

The policy documents construe risks and errors of AI not as products of human action but of AI technology, attributing error-creating agency to technology. This becomes apparent through the language employed. Although the need for ethical and legal frameworks raises concerns regarding AI being used for malicious ends, it is noticeable how passive language is still used: AI systems 'display intelligent behavior by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals'.[76] Similarly, it is AI technologies that 'require data to improve their performance',[77] not developers employing certain ML techniques; once 'they [AI technologies] perform well, they can help improve and automate decision making'.[78] Active language is used only when technology is seen to act, attributing a sort of agency to AI. AI systems are thus established as autonomous agents that need to be controlled by humans. At the same time, there is a sense of urgency that reflects technological determinism: these are societal problems that require legislative action immediately. This urgency indicates that this is a new situation, a tidal wave of digital transformation for which we need to brace ourselves, dictated by the inevitability of our novel historical situation. This sense of urgency resembles Ifversen's observation about the use of words such as 'crisis' or 'terrorism' to create necessity and legitimacy for policy action: 'such concepts defined for combat are used to frame a situation – or rather an event – that risks getting out of control'.[79] Because problems are portrayed as created by technology and not by economic and societal structures, AI is presented as a novelty and hence separate from the historical continuum of earlier forms of automation. The lack of acknowledgment of history also detaches policy dialogue from the lessons learned about the implementation of information systems in organizational decision-making in the 1960s and 1980s. This detachment makes it harder to contextualize AI problems in a meaningful way.

---

76 COM(2018) 237 final (n 53) 2.

77 Ibid.

78 Ibid.

79 Ifversen (n 20) 86.

At a textual level, the Guidelines create a contrast between human agency and the agency of AI systems, which again is established by the use of active language: 'AI systems ***should support*** human autonomy and decision making'; 'AI systems ***should both act as enablers*** to democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and ***allow for*** human oversight'.[80] Hence, there is a silence in the AI policy documents when it comes to humans. Only abstractions of humans are present, the human in control, but not the humans who could be perceived as objects of regulation. The abstraction of human control seems to imply that it is the closest human operator who is implicated in legal and ethical protection. Simultaneously, the technological focus ends up assigning subjectivity to AI and mystifying human capabilities as the last line of defense against this tide of foreign intelligence. In any case, attribution of agency to technology seems to be at the core of AI problematizations. It is the technology that needs to be subjugated.

In the Guidelines, human agency is presented as being under threat from AI systems and human oversight is proposed as the solution to protect this agency. The agency of humans is threatened by the agency of technology, signaling that the AI systems need to be forced to allow for human oversight to protect the latter's autonomy. Again, the human actors are missing. AI is anthropomorphized into an autonomous agent that might be malicious towards humans: AI systems 'may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy'.[81] It is this agency of technology that we need to be protected against and which poses a threat to fundamental rights and human agency, the latter of which is defined as the autonomy of the human user. Similarly, the Communication reflects an anthropomorphization of AI systems, locating the problem in the systems' ability to learn, which makes them autonomous from humans,[82] echoing the deeply-rooted concern for the loss of control. Humans remain out of sight, are hidden from sight: it is as if the AI systems develop independently to surpass humans, instead of programmers, systems architects and human

---

80 Guidelines (n 14) 15 [emphasis added].

81 Ibid 16.

82 COM(2019) 168 final (n 14) 2.

supervisors actively evaluating and implementing process automation by AI methods. But ultimately, the threat of technological agency to human autonomy is contained by human oversight. Despite the statement that oversight does not need to be total and continuous, there is no discussion on the implications of human oversight, whether and how human overseers are able to perform their oversight tasks nor what would be the criteria for human intervention or whether an overseer should possess some particular expertise.

3. Locating Silences: Promise of Control and Missing Humans

To summarize, the policy documents attribute autonomy and agency to AI systems and portray this agency as something that needs to be enabled and reined in at the same time. The attribution of agency to AI systems is particularly blatant in the use of active language: in the context of the problem representations, only AI systems are portrayed as actors. The technological agency is contrasted with threatened human values and rights. The threat, it seems, comes from the technology itself, not the developers nor implementing organizations. Humans remain absent except for the abstract formulations of human agency, autonomy and oversight that are the object and the subject of legal protection. Similarly, situations of shared control or hybridization of decision-making are not discussed, further emphasizing the juxtaposition of human and machines. In short, the EU policy documents reintroduce the human/machine dichotomy from the Fitts list into AI policy setting.

In addition, there is an apparent assumption that decision-making has been and still is a human exercise, but that this is about to change. The coming of AI is like a force of nature and the questions that remain are how to react to it and how to safeguard the fundamental values that are apparently in danger. It is assumed that AI systems are advanced enough to take independent action. The question is not raised as to what AI techniques can actually accomplish and what they cannot, nor why and how these techniques should be evaluated independently from other forms of automation. The fact that AI as such is not defined also raises the question of the extent to which these problematizations reflect the broader societal discussion on the perceived existential dangers of 'strong' general AI instead of narrow AI, which most

current applications are. At the same time, human understanding is perceived to be enough to rein in this powerful technology. In the end, we are absolved by the triumph of human cognitive skills, capable of seeing, understanding and acting on machine mistakes, demonstrating that these policy documents mystify the capabilities of human agency. Ultimately, humans trump machines.

There are several silences, most importantly the surprising absence of humans, despite the stated objective to place people at the center of AI development. Humans are present only as abstractions as the human in control or the human whose autonomy is at risk. These abstractions, however, detach control from context. Give that it is the context that defines applicable law and ensuing legal mechanisms to exercise control, it remains unclear how a human controller would be able to exercise meaningful oversight. In this sense, it is possible that the technological focus could lead to a regulatory standstill, due to the role of technological neutrality as a central legislative technique.

The focus on the importance of ethical standards prevents us from questioning the feasibility of soft law approaches rather than hard law regulation through binding accountability measures and market regulation. By assigning ethical concerns to technological agency, the discussion is framed in terms of the human/machine dichotomy. But do the ethical concerns described in particular in the Strategy, such as safety and liability, discrimination, cyber-attacks, not also exist outside the AI context? To what extent are the concerns, challenges and risks described related to particular AI techniques rather than broader trends of datafication, standardization, and automation? And, finally, how should we conceptualize and regulate situations of shared control over complex socio-technical systems, those in which human labor is inseparable from technological tools, in a way that does not unfairly assign responsibility to the closest human operators?[83]

As discussed above, the final step in the WPR analytical approach is self-reflexivity, which aims to make the analyzer's own problem representations explicit. I analyzed the three EU documents at a textual level, trying ascertain

[83] As Elish discusses, boundaries of agency become hard to pinpoint in situations of shared control over complex socio-technical systems. See Elish (n 16) 54.

whether or not the aforementioned academic discussions reflected in them. For example, was the shift from human/machine juxtaposition (the Fitts list) to problems of automation present in the policy documents and to what extent was the policy influenced by the interaction approach? Were the subjective ideological roots of human oversight normalized into objective knowledge? Where were the humans and technical systems at the textual level and which human agents were recognized as influencing the ethical and legal concerns of AI? Where do the vague ethical concerns emerge from and who creates them? These questions, although critical, take as a given the fact that human oversight often fails in repetitive monitoring tasks[84] and that the focus on human operators hides other human actors from sight. Based on the earlier research, the analysis built on the assumption that human oversight refers primarily to operators not systems designers, architects, and developers. However, the policy documents did not reflect this assumption, instead describing human agency only in abstract terms. On close critical reading, the absence of human actors becomes apparent and this silence further draws attention to the conspicuous use of passive language, which emphasizes the agency attributed to technology.

As my analysis was motivated by the concept of human control as a solution to AI problems and the observation of active/ passive language, I focused on relations between technical systems and human agents. Alternative approaches might focus on relations between different policy measures and their respective fields of law or discuss the economic agenda advocated particularly by the Commission. It should be noted, however, that inevitably there are other problem representations in these documents. As Bacchi and Goodwin note:

> it is highly likely that a WPR analysis may well need to be applied *more than once* in any particular applications. This is because problem representations tend to lodge or 'nest' one within the other.[85]

The need for repeated analysis of problematizations may become particularly vital when the EU's approach to AI governance is expanded from soft law guidelines to hard law instruments.

---

[84] See e.g. Bainbridge (n 29).

[85] Bacchi and Goodwin (n 50) 24.

IV. The Supremacy of the Human Overseer: Human Agency as Justification

Based on these observations, human agency, in the form of varying levels of human oversight, is portrayed as a central tool for overcoming the ethical concerns and risks associated with AI systems, regardless of the obvious limitations of human capabilities in monitoring automation. In the EU's AI policy, human decision-making is contrasted with algorithmic decision-making, thus invoking the human/machine dichotomy instead of collaboration in socio-technical systems. Through the process of juridification, the notion of human oversight becomes a legal concept, binding it to the internal rationality of law. This in turn limits the scope for critique: law only accepts critique when it is framed in terms law understands. Simply put, the dichotomy becomes embedded in legal doctrine and frames future socio-legal discussion.

The human control approaches reflect the assumed need to engage humans in algorithmic decision-making in order to ensure fairness and, ultimately, to address fears associated with automation and machines. Human oversight is about controlling unknowns, particularly unknowns of the technological variety. In this sense, human oversight as control over technology reflects something almost aspirational, a source of trust in the face of uncertainty. This promise of control is not simply about the feasibility of human oversight over automation, which has the obvious shortcomings discussed above. Could it be that human oversight carries this promise of control particularly because the fears and risks are attributed to technology, not the humans? This would suggest that humans are 'in the loop' to provide legitimacy, which is not necessarily linked to the practical feasibility of human oversight. Given this justificatory dimension of human agency, the feasibility of these approaches should be critically assessed before they are implemented as governance models, given that they may not in reality provide the solutions to the implicit problematizations. Instead, assigning problem-solving capabilities to human intervention might lead us astray as human oversight enables us to maintain law's 'human-facedness'.

Why then, despite the limitations of human control and its connection with the legal liability regime, do we still maintain the expectation that human input in decision-making is fundamental? It seems that this emphasis reveals

something relevant about law's self-reflection: a connection between human agency, legitimacy of decision-making, and social expectations of fairness.[86] The legal system produces 'human-faced' law, conceptualizing law in terms of human agents, which the problematizations around ADM reveal. However, by juxtaposing machines and humans we seem to imply that ADM systems are somehow fundamentally different decision-making mechanisms. But as the algorithmic bias discussion demonstrates, human subjectivity becomes embedded in ADM systems, making them, in many ways, as biased, arbitrary and subjective as human-driven decision-making albeit implemented on a different level. Do we still unconsciously expect our technological tools to be less subjective than we are? It seems that we often still assign objectivity to technology and feel betrayed when our expectations are not met.

Ultimately, however, human subjectivity wins against automation, as human oversight seems to imply. The strong preference towards human control, despite its limitations, suggests a deeper connection between human agency and the legitimacy of decision-making. In this sense, the policy documents use human control to justify the increase of automation. This reading echoes Elish's notion of humans as moral crumple zones and Jasanoff's pretensions of control, referred to above. In this sense, the emphasis on human control as the right policy solution for fundamental rights issues linked to ADM provides a particularly interesting viewpoint to the presumed socio-legal and regulatory challenges of AI.

There is a sense of urgency about AI presented in the policy documents, a call for action, which still boils down to voluntary soft law instead of hard law approaches. The emphasis on soft law may seem surprising, as the limitations of voluntary implementation are obvious. If the urgency portrayed in the guidelines is justified, why then only soft law? The chosen soft law strategy may be explained by the division of labor, as the President of the European

---

[86] Self-evidently conceptions of fairness are very much dependent on individuals, contexts, disciplines, fields, and theoretical backgrounds. On quantitative definitions see, Hutchinson and Mitchell (n 7); on human perceptions see e.g. Nina Grgic-Hlaca, Khrisna Gummadi, Elissa Redmiles, Adrian Weller, 'Human Perceptions of Fairness in Algorithmic Decisions Making: A Case Study of Criminal Risk Prediction' (Proceedings of the 2018 World Wide Web Conference. International World Wide Web Conferences Steering Committee, 2018) <https://dl.acm.org/citation.cfm?id=3186138> accessed 23 August 2019.

Commission, Ursula von der Leyen, intends to propose a comprehensive European approach to AI regulation in her first 100 days in office during spring 2020.[87] Nonetheless, the policy documents suggest that the problem with AI is not the lack of a legal framework as such but instead the more vague notion of an independent artificial agency. Perhaps the importance attributed to human input reflects the idea that legal decision making should not be about automated information processes but about processes that produce justification and legitimacy. Perhaps these policy statements come with a promise for renewed interest in the ritualistic elements and societal values present in legal decision making or, put another way, in conflict management, the production of justification through procedural structures.

Accountability mechanisms built on the assumption of a supreme human overseer are inherently flawed, if adopted without criticism. Such approaches can embed and reinforce the implicit human/machine dichotomy and mystify human agency. But it should be noted that the emphasis on human agency may serve a purpose outside monitoring automation, namely in justifying legal decisions. The importance attributed to humans in automation is not arbitrary but instead reflects the legal system's foundational concepts and ideologies that are built on anthropocentricity. In other words, juxtaposing algorithmic and human decision-making reveals law's self-reflection on what constitutes legal decision-making. Simply put, law's acknowledgement of legal agents capable of decisions is limited to humans or fictions of human agents such as organizations that are conceptualized as legal (although not natural) persons. Following this, justification of decision-making has traditionally been connected to human agency even when, in practice, decisions are arrived at through intra-organizational processes. In this sense, human control over automation can be seen simply as another formulation of human justification.

This analysis has attempted to demonstrate that problematizations do matter, perhaps more so in discussions concerning technological governance

---

[87] Ursula von der Leyen, 'A Union That Strives for More My Agenda for Europe' (2019) <https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf> accessed 27 November 2019.

than in other fields less plagued by misplaced metaphors.[88] It is not the objective of the WPR approach to provide alternative policy recommendations but instead to provide tools for critical analysis and to enable egalitarian politics.[89] By employing measures of critique, we are able to open the door to critical examination of automation of legal decision-making, the role played by human, non-human and hybrid actors in justification production, and ultimately, the feasibility of anthropocentric legal concepts to address this hybridisation. We can call attention to the justification of decisions and the legitimacy of decision-making processes and examine what exactly changes through implementation of ADM systems. Finally, the promise of this approach lies in a more nuanced understanding of how decisions come about. After acknowledging the fabricated nature of the problems associated with ADM, we can start thinking about meaningful partnerships between human agents and the automation of legal decision-making.

V. Conclusion: Implications for AI Policy and Socio-Legal Research

What implications follow from this analysis of the EU's AI policy focus on human control over automation? Two particular future avenues for analysis emerge. Firstly, what can and should we regulate and how can socio-legal scholarship facilitate the development of new effective regulatory strategies? As discussed, there is an inherent tension between technology-oriented policy and the principle of technological neutrality as guiding legislative strategy. This tension needs to be addressed if we want to pursue technological governance from an AI-specific perspective. The focus on technology may also be problematic due to the terminological ambiguity of AI and a significant theoretical issue relates to the legal system's limited focus on humans as objects of regulation. Difficult policy choices become entwined

---

88 See e.g. Sheldon Ungar, 'Misplaced Metaphor: A Critical Analysis of the "Knowledge Society" (2003) 40(3) Canadian Review of Sociology 331, 331-347; Marinus Ossewaarde, 'Digital Transformation and the Renewal of Social Theory: Unpacking the New Fraudulent Myths and Misplaced Metaphors' (2019) 146 Technological Forecasting and Social Change 24, 24-30.

89 Bacchi and Goodwin (n 50) 25.

with the need for critical socio-legal scholarship: should we forego the principle of technological neutrality or should we hold on to law's anthropocentricity? What exactly would these choices entail? If we stick with regulating human behavior, what criteria should be used to identify the 'human' in complex socio-technical systems? In any case, discussion of the objectives of AI regulation is unavoidable. To this end, a careful and context-specific analysis of the current legislative framework is needed to understand better whether existing legal safeguards possess enough interpretative flexibility to address the problems related to AI in the policy context. Such an examination also needs to address the efficiency of administrative and procedural safeguards beyond human control and contextualize the current debate within the broader historical development from the introduction of standards to data-driven automation of legal decision-making processes through information systems. If regulation is pursued, special attention should be paid to the creation of accountability mechanisms in a manner that does not impose unrealistic expectations on human operators but instead pursues a more rigorous interaction design. If we ignore the hybridization of legal decision-making that ADM models impose, there is a danger of assigning human decision makers the role of rubber-stampers with problematic consequences for legitimacy and justification.

Secondly, policy debates around AI ethics provide an interesting context in which to discuss the relationships between law, politics, and ethics, and reveals a way to examine the juridification of technological governance from soft law to hard law. As discussed, soft law guidelines are bound to frame the societal debate concerning AI challenges and their proposed solutions may have normative consequences in shaping future hard law instruments. The juridification of such concepts, i.e. their translation into binding concepts of positive law, also disguises the heterogenous value-laden and ideological choices present in the political creation of AI ethics guidelines. The juridification binds concepts established in policy-making to the legal sphere's internal perspective. This process reflects the long-standing distinction between the political and legal systems and their separate societal functions, built on the idea that the political system debates societal objectives and establishes a compromise in the form of legislation, after which the legal system takes over its application and interpretation. This means that within the legal system there is limited space for fundamental

critique about the acceptability of legislative intent. In other words, the translation from politics to law limits the grounds on which these concepts can be criticized: from the legal system's normative view, only immanent critique preserving law's internal rationality is recognized as valid.[90]

What makes human control over automation such a tempting solution for digital technologies is its relatively easy implementation. Human control comes with the promise of a relatively simple and operational way to address AI-related issues: to operationalize human control both within the legal system and software development requires relatively easy political choices that can be met by establishing a legal right to human oversight and then creating technical design solutions for implementing this right within the technological architecture. Especially if the alternative is to engage in grueling societal debates over the dynamics of technological change and the critical analysis of existing societal power imbalances, such solutions provide attractive and straightforward policy actions.[91] But perhaps the latter is exactly what is needed. To better understand the complexities and societal issues related to the ongoing algorithmization and to assess different policy options, it is necessary to broaden the discussion from the current focus on technology and ethics to discussions about societal structures and law.

---

[90] On immanent critique, see Kaarlo Tuori, *Critical Legal Positivism* (Ashgate 2002) 29–30. See also Riikka Koulu, *Law, Technology, Dispute Resolution: Privatisation of Coercion* (Routledge 2019) 37.

[91] Cf. Kenneth C. Laudon, *Computers and Bureaucratic Reform: The Political Functions of Urban Information Systems* (John Wiley & Sons 1974) 52-53.